Learning realistic human actions from movies

Ivan Laptev
INRIA Rennes,IRISA
ivan.laptev@inria.fr

Marcin Marszalek
INRIA Grenoble,LEAR-LJK
marcin.marszalek@inria.fr

Cordelia Schmid
INRIA Grenoble,LEAR-LJK
cordelia.schmid@inria.fr

Benjamin Rozanfeld
Bar-Ilan University
grurgrur@gmail.com

Abstract

这篇文章的目的是处理多样化以及现实生活的视频场景设定中的人类行为识别问题。这项困难但是重要的课题在过去因为诸如缺乏现实标记的视频数据集等问题被忽视。我们的第一项工作是处理数据集的限制，研究使用电影剧本来自动标注视频中的人类行为。同时我们评估了其它可选的行为恢复方法进行比较并展示了基于文本的分类器的优势。采用恢复出来的行为样本进行视觉学习，然后我们继续处理视频中行为分类的问题。我们提出了一种新的基于视频的分类方法，这些方法依赖和扩展于一些近来提出的方法，包括局部空域 -时域特征，空域 -时域金字塔以及多通道非线性支持向量机。我们提出的方法与其它具有代表性的技术相比在标准 KTH 行为数据集上将精确度提升到了 91.8%，而且我们特别研究并展现了我们的方法对于训练数据集标注错误的非常高的耐受性。最后，我们将方法应用到电影中具有挑战性的行为类别上，呈现出可观的结果。

Figure 1. 三类人类行为的现实样本：接吻，接电话，走出汽车。所有样本都通过从剧本对齐的电影中自动采集得到。

1. 引言

在过去的十年中视觉识别领域从分类玩具对象实例发展到识别自然图像中多种类的对象和场景，有了显著的进步。这受益于现实的图像数据集的创建以及新的鲁棒的图像描述和分类方法。我们从这一进程中受到启发，计划将先前的经验转换到视频识别和人类行为识别的领域中。

已有的人类行为识别数据集 (例如 [14]，见图 8) 只提供了处于控制和简化的场景设定下记录的较少的行为类别。这和现实生活应用要求的处理包含具有个体差异的人类行为的自然视频有很大差异，这些个体差异来源于表情，姿势，动作和衣着，透视效果和镜头运动，明亮度差异，以及场景遮挡和变化等。这篇文章中我们将处理当前数据集的限制，采集现实视频中的人类行为样本 (如图 1 所示)。特别地，我们考察人工视频标注的困难并提出自动标注电影中的人类行为的基于剧本对齐和文本分类的方法 (第 2 节)。

视频行为识别和静态图片行为识别面临着同样的困难。都需要处理明显的类内差异，背景混杂和遮挡问题。在静态图片对象识别中，这些问题能够通过特征包 (bag-of-features) 描述方法 [16]和例如 SVM 这样的具有代表性的机器学习技术组合起来得到相当好的处理。但是，这些结果是否能推广到现实中的人类行为识别中尚有待解答，例如，对故事片或者个人录像进行识别。

基于近来图像分类的经验，我们在时空域上采用时空特征和广义的空域金字塔。这使得我们可以用弱几何信息扩展时空特征包描述算法，并应用基于运算核的学习技术 (第 3 节)。我们在标准的基准 [14]上进行验证，发现这样优于其它具有代表性的技术。然后我们转向现实视频中的行为分类问题，并展示了我们采用的方法在电影中 8 类非常具有挑战性的行为类别上的可观结果。最后，我们提出并评估一种行为学习和分类的全自动化配置，获取自动化标注的数据集合。

1.1. 相关工作

我们的基于剧本的人类行为标注技术在本质上和一些近来采用文本信息从网络上 [6, 13]进行自动图像采集以及给图像中 [1]和视频中 [4] 的角色自动命

名的文章相似。不同的是这项工作采用更加精细的文本分类工具来克服文本中的行为差异。类似的，一些近期的方法探索采用特征包描述法来进行行为识别[3, 5, 12, 14, 17]，但只是针对于控制下的简化场景设定。电影中的行为的识别和定位近期在 [8]中在有限的数据集上得到处理，即人工标注的两类行为。这里我们提出一种适用于数十个甚至更多的视觉行为类的自动标注。我们的视频分类方法借鉴了一些图像识别方法 [2, 9, 11, 18]，并将空域金字塔 [9]扩展到时空域金字塔。

2. 人类行为自动化标注

这一节描述从电影中采集带标注人类行为的视频数据的自动过程。电影里有非常多样的和大量的现实人类行为。然而，诸如接吻，接电话以及走出汽车等常见的行为类 (图 1) 在一部电影中只出现很少的次数。为了获取充足数量的行为样本以用于视觉训练，对上百个小时的视频进行标注是必要的，人工标注是非常困难的一项任务。

为了避免人工标注的困难，我们采用电影剧本 (或简称 “剧本”)。上百部著名的电影[1]的剧本是公开的并且在场景，角色，转录对白和人类行为等方面提供了电影内容的描述。将剧本作为视频标注的方式已被 Everingham 等人 [4]用于视频中的角色的自动命名。这里我们扩展这种思想，并应用基于文本的剧本搜索来自动地收集人类行为的视频样本。

然而，从剧本中自动标注人类行为同样面临许多问题。首先，剧本通常没有时间信息，所以必须要和视频进行对齐。其次，剧本中描述的行为并不是常常都和电影中的行为相关。最后，行为恢复必须要处理文本中行为的大量实质性的变化。在这一节中我们在 2.1 节和 2.2 节中处理这些问题，并在 2.3 节中采用提出的方法采集标注了人类行为的视频样本。得到的结果作为第 4 节中训练和评估视觉行为分类器的数据集。

2.1. 剧本和视频行为的对齐

电影剧本通常是普通文本格式并具有相似的结构。我们利用行缩进作为简单特征来将剧本解析为独白，角色名称和场景描述 (图 2)。为了将剧本和视频对齐我们采用 [4]的方法并利用另外从互联网上下载的电影字幕中的时间信息。如同 [4]一样，我们首先采用单词匹配和动态编程对齐剧本和字幕中的对话段。然后将字幕中的时间信息转移到剧本中，并推测场景描述之间的时间间隔，如图 2 所示。本文用于行为训练和分类的视频剪辑采用场景描述之间的时间间隔来定义，可能包含多个行为或是没有行为的片段。为了指出一个由剧本和字幕的不匹配造成的可能的不对齐的情况，我们将每一个场景描述和分数 a 关联起来。a 通过匹配字数与邻接对白的比例计算：$a = (\#matchedwords)/(\#allwords)$。(译者注：‘#’ means “number of”。)

[1]我 们 从www.dailyscript.com, www.movie-page.com以及www.weeklyscript.com上得到了上百个电影剧本。

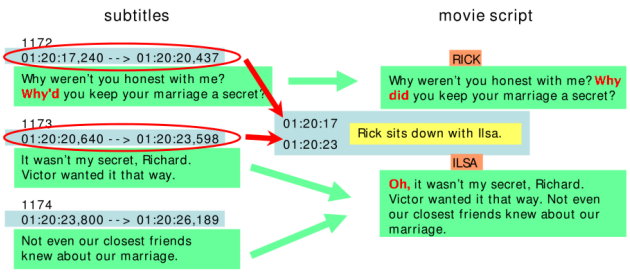

Figure 2. 字幕和剧本中对话片段（绿色）的匹配示例。相邻对话片段的时间信息（蓝色）用来估计场景描述（黄色）的时间间隔。

时域不对齐可能由剧本和字幕不符造成。然而，由于可能的剧本和电影的不符，完美的字幕对齐 ($a = 1$) 并不能保证对视频行为的正确标注。

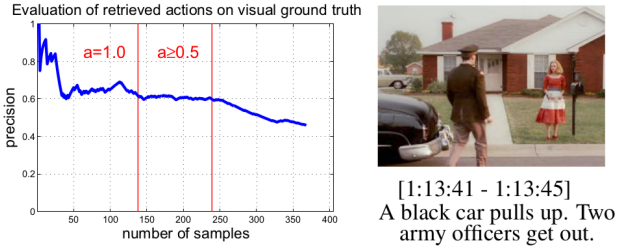

Figure 3. 基于剧本的行为标注的评估。左：基于真实信息的行为标注精确度。右：“走出汽车 “ 这一行为的视觉阳性误判的示例。

为了处理这个问题，我们人工地标注 12 部电影剧本中的上百个行为并在视觉真实情况上验证。147 个具有正确文本对齐 ($a = 1$) 的行为中只有 70% 和视频匹配。不匹配的样本可能是时间不对齐 (10%)，在视野之外 (10%) 或是完全不在视频中 (10%)。字幕不对齐 ($a < 1$) 进一步降低了视觉精确度，如图 3 左所示。图 3 右展示了由于不在摄像头视野中所造成的 “走出汽车” 这一行为的一种典型的 “视觉阳性误判”。

2.2. 基于文本的人类行为恢复

文本中的人类行为表达方式可能会有一些类内的差异。下面的例子描述了 “走出汽车” 这一行为的表达方式的一些类内差异：“Will gets out of the Chevrolet.”，“A black car pulls up.Two army officers get out.”，“Erin exits her new truck.”。而且，阳性误判不容易从阳性样本中分离出来，如 “坐下” 这一行为：“About to sit down,he freezes.”，“Smiling,he turns to sit down.But the smile dies on his face when he finds his place occupied by Ellie.”。因此，基于文本的行为恢复并不是想象那么简单，通过像 [13]那样为了恢复图像中的对象而采用简单的关键字搜索是很难解决问题的。

为了处理人类行为文本描述方式的类内变化，我们

采用基于机器学习的文本分类方法 [15]。分类器对剧本中的每个场景描述进行标记 - 包含目标行为或没有包含目标行为。实现方法依赖于特征包模型，每个场景描述用一个高维特征空间中的稀疏向量表示。至于特征我们采用一个 N 个单词的小窗口 (N 在 2 到 8 之间) 中的单词，邻接单词对，以及非邻接单词对。少于 3 个训练文档支持的向量被删除掉。分类方面我们采用一个和支持向量机等价的正则感知器 [19]。分类器在人工标注的场景描述集合上训练，参数 (正则常数，窗口尺寸 N，以及判决门限) 使用验证集合进行调整。

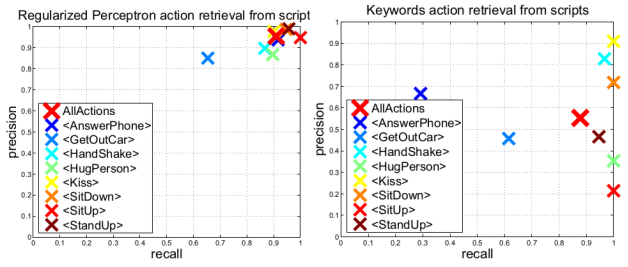

Figure 4. 采用正则感知分类器 (左) 和正则表达式匹配 (右) 进行 8 类人类行为类恢复的结果对比。

我们在本文一直采用的 8 类电影中的行为上对基于文本的行为恢复进行评估：接电话，走出汽车，握手，拥抱，接吻，坐下去，坐起来，立正。文本测试集合包含 12 个人工标准的电影剧本里的 397 个行为样本以及超过 17000 个非行为样本。文本训练集合从不同于测试集合的大量的剧本里采样。我们比较正则感知分类器和通过匹配文本中的人类行为人工调整的正则表达式得到的结果。图 4 中的结果明显展示了文本分类器的优势。平均 precision-recall 值分别是文本分类器 [prec. 0.95/rec. 0.91] 和正则表达式匹配 [prec. 0.55/rec. 0.88]。

2.3. 人类行为视频数据集

我们构建了两个视频训练集合 (一个人工标注，一个自动标注)，以及一个视频测试集合。这些视频包含了 8 类电影中行为 (图 10 第一行) 的视频剪辑。在每种情况中我们首先使用自动剧本对齐技术 (2.1 节)。对于人工数据集和测试集合我们人工地从人工文本标注行为的剧本中选择了视觉上正确的样本。自动化数据集包含了用 2.2 节描述的文本分类器自动恢复出来的样本。我们限定自动化训练集合为包含 $a > 0.5$ 的行为，并将视频长度限定在 1000 帧以内。人工和自动化的训练集合包含了 12 部电影[2]中的视频序列，测试集合来自另外 20 部电影[3]。我们的数据集 (视频剪辑及相应的标注) 可从http://www.irisa.fr/vista/actions下载。

采用两个训练数据集的目的是对在监督学习设定和自动生成训练样本设定下的设别进行评估。注意到自

[2]“American Beauty”,“Being John Malkovich”,“Big Fish”, “Casablanca”,“The Crying Game”,“Double Indemnity”,“Forrest Gump”,“The Godfather”,“I Am Sam”,“Independence Day”, “Pulp Fiction” and “Raising Arizona”

[3]“As Good As It Gets”,“Big Lebowski”, “Bringing Out The Dead”,“The Butterfly Effect”,“Dead Poets Society”,“Erin Brockovich”,“Fargo”,“Gandhi”, “The Graduate”,“Indiana Jones And The Last Crusade”,“Its A Wonderful Life”,“Kids”,“LA Confidential”,“The Lord of The Ri ngs:Fellowship of the Ring”,“Lost Highway”,“The Lost Weekend”,“Mission To Mars”,“Naked City”,“The Pianist” and “Reservoir Dogs”.

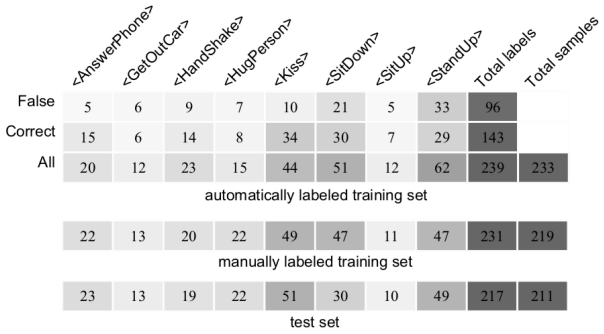

Table 1. 自动训练集合（上），人工训练集合（中）以及测试集合（下）中的行为标签数量。

动化训练集合中的样本和视频都没用采用任何人工标注。不同子集的行为标签和行为类的分布在表 1 中给出。我们观察到正确标注视频的数量在自动化集合中占 60%。大多数错误标注来自于剧本和视频不对齐以及少部分来自于文本分类器的附加错误。分类导致的错误训练标注的问题将在 4.3 节中进行讨论。

3. 行为识别中的视频分类

这一节描述我们采用的行为分类方法。方法基于已存在的用于视频描述的特征包方法 [3, 12, 14]并将静态图片分类中的优势扩展到视频中 [2, 9, 11]。Lazabnik 等人 [9]发现空域金字塔，即空域场景布局的粗描述法，能够提升识别性能。基于这种方法的一些成功的扩展包括单层金字塔的权重最优化 [2]以及广义空域网格 [11]的采用。这里我们基于这些想法提出构建空域 -时域网格的方法。方法的具体细节在下文描述。

3.1. 空域 -时域特征

稀疏空域 -时域特征近来表现出在行为识别方面的良好性能 [3, 5, 12, 14]。这些方法提供了一种紧凑的视频描述并且对背景混杂，遮挡和尺寸变化具有耐受性。这里我们采用 [7]，使用 Harris 操作的空域 -时域扩展检测兴趣点。然而，和 [7]中采用缩放选择不同，我们使用多尺度方法并在多层空域 -时域尺寸 (σ_i^2, τ_j^2) 上进行特征提取，$\sigma_i = 2^{(1+i)/2}, i = 1, ..., 6; \tau_j = 2^{j/2}, j = 1, 2$。这种选择受启发于减少计算复杂度，独立于尺寸选择人工痕迹以及近期证明的采用密集尺寸缩放采样的良好识别性能。我们同时消除由于人工痕迹造成的位于判决边界的检测 [10]。图 5 展示了两帧人类行为图像中检测出的兴趣点。

Figure 5. 两帧带有人类行为：握手（左），走出汽车（右）中检测出的空域 -时域兴趣点

为了使局部特征的动作和外貌特征化，我们计算邻接检测点空域 -时域卷的直方图描述符。每一卷的大小 $(\delta_x, \delta_y, \delta_z)$ 和检测尺寸相关，$\delta_x, \delta_y = 2k\sigma, \delta_t = 2k\tau$。每一卷被分为一个 (n_x, n_y, n_t) 的立方体网格；对每一个立方体我们分别计算粗梯度方向直方图 (HoG) 和光流直方图 (HoF)。标准化后的直方图通过多列索引串联为 HoG 和 HoF 描述符矢量，这本质上和著名的 SIFT 描述符相似。参数方面我们选择 $k = 9, n_x, n_y = 3, n_t = 2$。

3.2. 空域 -时域特征包

给定一个空域 -时域特征集合，我们构建一个空域 -时域特征包 (BoF)。这需要构建一个视觉单词。在我们的实验中我们对一个从训练视频中提取的 100,000 个特征子集采用 k-means 算法进行聚类。聚类数量设定为 $k = 4000$，经验显示这个参数能够给出良好结果，这和静态图像分类是一致的。然后 BoF 表达式将每一个特征分配给最近的 (欧式距离) 单词，并在空域 -时域卷上计算视觉单词出现次数直方图，空域 -时域卷与整个视频序列和空域 -时域网格定义的子序列相关。如果有多个子集，不同的直方图串联成一个特征矢量，然后进行标准化。

空域维度方面我们采用一个 1×1 网格 -与标准 BoF 表达式相关，以及一个 2×2 网格 -在 [9]中展示出了优异的结果，一个水平 $h3 \times 1$ 网格 [11]以及一个垂直 $v1 \times 3$ 网格。同时，我们实现一个密集 3×3 网格以及一个中心聚焦 $o2 \times 2$ 网格，邻接单元 50% 重叠。时域维度方面，我们将视频序列分为 1 到 3 非重叠时域箱，即 t_1,t_2 及 t_3 分箱。t_1 表示标准 BoF 方法。我们也实现了一个中心聚焦的 ot_2 分箱。注意到重叠部分的网格位于中心的特征权重更大。

6 个空域网格和 4 个时域分箱的组合形成 24 种可能的空域 -时域网格。图 6 描述了一些被证明是对行为识别有用的网格。每一个空域 -时域网格的组合使用一个 HoG 或 HoF 描述符，封装进下面介绍的一个通道中。

3.3. Non-linear Support Vector Machines

为了进行分类，我们采用非线性支持向量机和一个鲁棒地联合通道的多通道 χ^2 核 [18]。采用多通道高斯核：

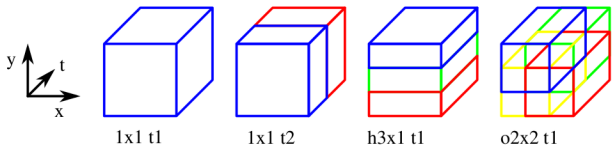

Figure 6. 一些空域 -时域网格的示例。

$$K(H_i, H_j) = exp(-\sum_{c \in C} \frac{1}{A_c} D_c(H_i, H_j))$$

$H_i = h_{in}; H_j = h_{jn}$ 是通道 c 的直方图，$D_c(H_i, H_j)$ 是 χ^2 距离：

$$D_c(H_i, H_j) = \frac{1}{2} \sum_{n=1}^{V} \frac{(h_{in} - h_{jn})^2}{h_{in} + h_{jn}}$$

V 是单词大小。参数 A_c 是一个通道的所有的训练样本之间的距离的平均值 [18]。针对给定的训练集合的 C 通道的最佳集合通过贪心算法找到。从空通道集合开始所有的可能的附加核删除通道都被评估，直到达到最大值。在多类分类方面我们使用 one-against-all 方法。

4. Experimental results

4.1. Evaluation of spatio-temporal grids

4.2. Comparison to the state-of-the-art

4.3. Robustness to noise in the training data

4.4. Action recognition in real-world videos

5. Conclusion

References

[1] T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. Forsyth. Names and faces in the news. In Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, volume 2, pages II–848–II–854 Vol.2, June 2004. 1

[2] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In ACM International Conference on Image and Video Retrieval, 2007. 2, 3

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on, pages 65–72, Oct 2005. 2, 3

[4] M. Everingham, J. Sivic, and A. Zisserman. "Hello! My name is... Buffy" – automatic naming of characters in TV video. In Proceedings of the British Machine Vision Conference, 2006. 1, 2

[5] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, pages 1–8, Oct 2007. 2, 3

[6] L. jia Li, G. Wang, and L. Fei-fei. Optimol: automatic online picture collection via incremental model learning. In In CVPR, 2007. 1

[7] I. Laptev. On space-time interest points. Int. J. Comput. Vision, 64(2-3):107–123, Sept. 2005. 3

[8] I. Laptev and P. Perez. Retrieving actions in movies. In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, pages 1–8, Oct 2007. 2

[9] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, volume 2, pages 2169–2178, 2006. 2, 3, 4

[10] R. Lienhart. Reliable transition detection in videos: A survey and practitionerâ s guide. International Journal of Image and Graphics, 1:469–486, 2001. 3

[11] M. Marszałek, C. Schmid, H. Harzallah, and J. van de Weijer. Learning object representations for visual object class recognition, oct 2007. Visual Recognition Challange workshop, in conjunction with ICCV. 2, 3, 4

[12] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. Int. J. Comput. Vision, 79(3):299–318, Sept. 2008. 2, 3

[13] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, pages 1–8, Oct 2007. 1, 2

[14] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, volume 3, pages 32–36 Vol.3, Aug 2004. 1, 2, 3

[15] F. Sebastiani. Machine learning in automated text categorization. ACM Comput. Surv., 34(1):1–47, Mar. 2002. 3

[16] J. Willamowski, D. Arregui, G. Csurka, C. R. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In In ICPR Workshop on Learning for Adaptable Visual Systems, 2004. 1

[17] K.-Y. K. Wong, T.-K. Kim, and R. Cipolla. Learning motion categories using both semantic and structural information. In Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, pages 1–6, June 2007. 2

[18] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. In Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06. Conference on, pages 13–13, June 2006. 2, 4

[19] T. Zhang. Large margin winnow methods for text categorization, 2000. 3